****

**The Battle of the Neighbourhoods – Opening a Clothing Store in London, UK**

**IBM Applied Data Science Capstone Project**

**By: Sakib Chughtai**

**Date: 2/12/2020**

**Table of Contents**

1. **Introduction**
   1. **Background**
   2. **Business Understanding**
   3. **Analytics Approach**
2. **Data**
   1. **Data Sources**
   2. **Data Cleaning**
   3. **Feature Engineering**
3. **Exploratory Data Analysis**
4. **Predictive Modelling**
5. **Conclusion**
6. **Future Work**
7. **References**
8. **Appendix**

**1 Introduction**

* 1. **Background**

With over 8.5 million inhabitants, London is one of the largest and most influential cities in the world. Often regarded as one of the world's most connected cities, London has a diverse range of cultures with over 250 languages spoken in the region. Each year, the city attracts over 21 million international visitors as well as 28 million domestic tourists.

London is home to over 40,000 shops and 26 major street markets. Oxford Street is known as London’s busiest shopping street. Located in the heart of the city’s West End, Oxford Street offers a wide variety of clothing stores from high end luxury brands such as Louis Vuitton, Burberry and Gucci to high street fashion brands such as Zara, Tommy Hilfiger and Polo Ralph Lauren.

There are many other types of clothing stores located in London to suit the different needs of customers, such as The North Face and Patagonia which sell mountain and winter clothing. There are also cheaper options for people looking to buy clothes on a budget, such as Primark and H&M. All in all, there is a clothing store for anyone and everyone.

* 1. **Business Understanding**

The goal of this project is to predict the optimal location to open a clothing store in London, UK. This report will provide stakeholders with an insight into which areas are the most promising based on a multitude of factors.

Tackling this problem head on requires breaking down the end goal into objectives that are in support of the goal. By breaking down the objectives, structured discussions can take place where priorities can be identified in a way that can lead to organizing and planning how to tackle the problem at hand.

There are thousands of different clothing stores in London. The first objective is to locate all of these stores using the Foursquare API. We will also look at all shopping malls because this may have an impact on where the ideal location to open a store will be. The next step consists of using demographic data such as population density, average age and income to cluster similar areas in London. The optimal location to open a store will be in an area where there is a high concentration of clothing stores AND shopping malls. Ideally, this area should also have a high population density and average income.

* 1. **Analytics Approach**

Selecting the right analytic approach is key to ensuring we answer the question being asked correctly and accurately. Given that we are trying to show relationships between different features within the model, we will use a descriptive model. We will use K-means clustering which is a simple and popular unsupervised machine learning algorithm to answer our question. A cluster is defined as a collection of data points that are aggregated together because of certain similarities. This will be the best technique to use for our use case.

We will use several data science tools to achieve the end goal. First, we must collect the required datasets, clean them and then merge them into one big dataset to prepare for the modelling stage. We will then use machine learning and visualization tools to generate the most promising areas based upon criteria we defined earlier.

Additionally, we will evaluate the quality of the model by through a diagnostic measures phase and statistical significance testing. These evaluation techniques ensure the model is working as intended and that the data is being properly handed and interpreted. Lastly, we will present our findings to the stakeholders so they can make an informed decision on where they would like to open their clothing store

1. **Data**
   1. **Data Sources**

If the problem that needs to be solved is a ‘recipe’, the data is an ‘ingredient’, so to speak. As a data scientist, there are several aspects of the data that need to be addressed before moving onto the data preparation stage.

We need to identify what datasets are needed, how to source or collect them, how to understand or work with them and how to prepare the data to meet the desired outcome. Based upon the criteria we defined earlier, there are several factors that will impact the final decision of which area to open a clothing store. These are:

* The number of clothing stores within the neighbourhood
* The number of shopping malls within the neighbourhood
* London census information – population, density, average age, average income

We will define each neighbourhood using a grid format around the whole city. This grid will consist of many hexagons in a honeycomb layout. This will be the basis for defining different boroughs and neighbourhoods.

We will collect the data from the following sources:

* London borough boundaries shapefile which is defined using longitude and latitude values. Will be in geojson format to be easily integrated into analysis. Can be obtained from the London Datastore (publicly available).
* Coordinates which define the centre of London, this will be obtained using the **Google Geocoding API**.
* Clothing stores and shopping mall data in every borough will be obtained using **Foursquare API.**
* The centre of every borough will be defined using hexagon cells in a honeycomb layout, these will be calculated algorithmically. The addresses for those centres will be fetched by using the **Google Geocoding API**.
* London census data for the year 2016 will be obtained from the London Datastore (publicly available).
  1. **Feature Engineering**
  2. **Data**